Granularity-tunable Gradients Partition (GGP) Descriptors for Human Detection

Yazhou Liu[1], Shiguang Shan[2], Wenchao Zhang[1], Xilin Chen[2], Wen Gao[3,1]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing,China
[3]Institute of Digital Media, Peking University, Beijing, China
{yzliu, sgshan, wczhang, xlchen, wgao}@jdl.ac.cn

Abstract

This paper proposes a novel descriptor, granularity-tunable gradients partition (GGP), for human detection. The concept granularity is used to define the spatial and angular uncertainty of the line segments in the Hough space. Then this uncertainty is backprojected into the image space by orientation-space partitioning to achieve efficient implementation. By changing the granularity parameter, the level of uncertainty can be controlled quantitatively. Therefore a family of descriptors with versatile representation property can be generated. Specifically, the finely granular GGP descriptors can represent the specific geometry information of the object (the same as Edgelet); while the coarsely granular GGP descriptors can provide the statistical representation of the object (the same as histograms of oriented gradients, HOG). Moreover, the position, orientation, strength and distribution of the gradients are embedded into a unified descriptor to further improve the GGP's representation power. A cascade structured classifier is built by boosting the linear regression functions. Experimental results on INRIA dataset show that the proposed method achieves comparable results to those of the state-of-the-art methods.

1. Introduction

The research of human detection has received more attention in the recent years because of increasing demands in practical applications, such as smart surveillance system, on-board driving assistance system and content based image/video management system. Even through remarkable progress has been achieved [2, 15], finding the human in the still image is still one of the hardest tasks for object detection. The difficulties come from the human body's articulation, the occlusion and clothes/illumination variation.

A variety of features have been invented to overcome the difficulties mentioned above. Earlier works for human detection started from Haar-like features, which have been applied to face detection task successfully [16]. Papageorgiou et al.'s [14] used dense Haar-like feature as the basic descriptor for the human body and used the Support vector machine (SVM) as the classifier. Mohan et al. [13] extended this work by representing the human body as the combination of several body parts. Viola and Jones [17] showed that the combination of static and dynamic information by Haar-like features can be an effective way for finding the moving human, especially when the image size of the human is small.

Because of the large variation of the human clothes and the background, some researchers turned to the gradient based descriptors. One main category of gradient based method is to use shape/contour to represent human body. Gavrila [6] presented a contour based hierarchical chamfer matching detector. Lin et al. [10] extended this work by decomposing the global shape models into parts and con-

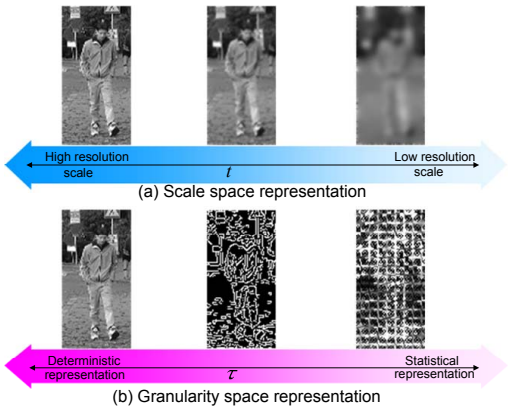

Figure 1. The difference between scale-space and granularity-space; scale-space is to represent the image with multiple resolutions; the granularity-space is to represent the image with multiple-level of summarizations, ranging from deterministic representation to statistical representation

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

structing a hierarchical tree for the part templates. Ferrari el al. [5] used the network of contour segments to represent the shape of the object. Wu and Nevatia [18] proposed edgelet to represent the local silhouette of the human. Another main category is to use the statistical summarization of the gradients. E.g., in [11], Lowe introduced the famous SIFT descriptor for object recognition. Mikolajczyk et al. [12] introduced position-orientation histogram features. Leibe et al. incorporated the SIFT descriptor into their implicit shape model(ISM) for human detection in [9] and extended this work into a street scene reconstruction system in [8].

One of the breakthroughs for human detection comes from Dalal and Triggs' work in [2]. They proposed histograms of oriented gradients (HOG) as the descriptor and used SVM as the classifier. Their results showed that the HOG can outperform the previous methods by a big margin on INRIA dataset. In [3], they extended this work into the space-time domain to detect the moving human from the video sequence. Zhu et al. [21] extended Dalal's work by combining HOG with a cascade structured detector.

After HOG, another big advance comes from Tuzel el al.'s work in [15]. They utilized the covariance matrix as the descriptor for human representation. Their method can encode the gradients' strength, orientation and position information in symmetric positive definite covariance descriptors which lie on a Riemannian manifold. Comparing with the best results of HOG, this method can further reduce the miss rate by 2.5 percents at 10^{-4} FPPW.

Besides the invention of the new features, some works show that using the combination of existing features can also improve the performance. Han el al. [7] used generalized Swendsen-Wang cut to generate the composite of Haar-like features and their results showed that this composition leads to generic improvement for the Haar-like features. More recent results in [4] showed multiple instance learning can be a possible solution for the misalignment for human body.

Wu and Nevatia[19] combined the existing heterogeneous features, e.g. edgelet, HOG and covariance matrix, into a boosting framework to improve both the accuracy and the speed. The advantage of this combination seems straightforward, since the heterogeneous features can "see" different things. For example, the HOG can be considered as a kind of *statistical* summary of the object. It is robust to noise and articulation but not suitable for describing the accurate structures. Edgelet descriptor can provide the *deterministic* representation of the specific shape and contour, but it is less robust to the rotation and translation variation. So the description abilities of these descriptors are complementary and the combination of them should yields better description ability.

A question arises here naturally is that: is there a unified framework to accommodate these seemingly heterogeneous descriptors, from deterministic to the statistical? If this framework can be found, then on the one hand, instead of merely "deterministic" or "statistical", some "gray-region" descriptors can also be deduced, which can further enrich the features' representation ability; on the other hand, under the same framework, the complexity of the algorithm can be reduced dramatically. This is the motivation of the work in this paper.

Here, we use the concept of *granularity* to represent the feature's description properties that range from *deterministic* description to *statistical* representation. Deterministic description means that the features can represent the specific structures and shapes, such as edgelet; while the statistical representation means that the features can only provide the statistical summarizations of the shapes and structures, such as HOG. The features of different granularity can generate a family of representations of the object, which we referred to as *granularity space representation*. Ideally, this granularity space can be parameterized by a parameter τ, which we referred to as the granularity parameter (or granularity for short).

The definition of the *granularity space* above is parallel to the definition of the scale space. So it is worth noting the relation and difference between these two concepts. Both of them intend to build multiple representations of the objects in the real world. For the scale space, this multiple representations come from different resolutions, which mimic the from-sharp-to-blur observation of an object at different distances. But for the granularity space, the multiple representations come from different level of summarization, ranging from the deterministic description to statistical representation. Please refer to Fig.1 for an intuitive illustration.

The granularity space is more or less in line with Wu et al.'s work [20], in which the authors indicated that the visual space can be partitioned into different regimes according to the entropy rate and the objects appearing at different entropy regimes will present different statistical properties. So, for the image reconstruction task, they use sparse coding (deterministic description) for low entropy regimes reconstruction and Markov random fields (statistical representation) for high entropy regimes reconstruction. Instead of using two different features to represent different statistical properties, this paper is trying to develop one feature which can adapt to different statistical properties to represent the varying characteristics of the complex objects.

The rest of the paper is organized as follows: Section 2 gives the outline and formulation of the propose method; Section 3 provides the implementation details; and Section 4 presents the experimental results.

2. Granularity-tunable gradients partition (GGP) descriptors

In this section, we describe the mathematical definition of the granularity. More intuitively, the granularity is used to represent the uncertainty of line segments in the Hough space. Fine-granularity corresponds to the line segments with low uncertainty. The geometry properties of these lines are clear and specific, and they can provide the deterministic representation of the object. Coarse-granularity corresponds to the line segments with high uncertainty. The geometry properties of these lines are weak and statistical information become dominant, and they can provide statistical representation of the object.

Formally, given an image I, we represent it by a family of descriptors with different granularities as

$$\left\{\bar{d}_{\vartheta,\tau} | \bar{d}_{\vartheta,\tau} = f(I;\vartheta,\tau), \tau \in \Gamma, \bar{d}_{\vartheta,\tau} \in \Pi_\tau\right\} \quad (1)$$

where the function $f : I \mapsto \Pi_\tau$ is a mapping from the original image space I to the feature space Π_τ, τ is the granularity parameter and ϑ is the feature parameter. More intuitively, $f(\cdot)$ can be considered as a feature extraction function. For this function, the input is image I; the feature is specified by the parameter ϑ; the granularity is specified by the parameter τ; and the output $\bar{d}_{\vartheta,\tau}$ is a descriptor of the image I. The Equ.1 is a general formulation of the granularity space representation. Under this framework, the feature extraction process is unified and the granularity property of the output feature can be controlled by an input parameter.

We developed our GGP descriptor under the framework described by Equ.1, and specialized the mapping function as

$$f(I;\vartheta,\tau) = S\left(T(I;\vartheta,\tau)\right) \quad (2)$$

where the feature extraction function $f(\cdot)$ is represented by the composition of two functions $T(\cdot)$ and $S(\cdot)$. The $T(\cdot)$ is referred to as image parsing (IP) function and $S(\cdot)$ is referred to as image description (ID) function. Intuitively, our GGP feature extraction procedure contains two steps: firstly, we parse the original image into the structure primitives by $T(\cdot)$; and secondly, we generate the descriptions of structure primitives by $S(\cdot)$. The granularity control is accomplished by image parsing function.

2.1. Image parsing

The structure primitives used in GGP are line segments. And the image parsing function $T(I;\vartheta,\tau)$ can be considered as a generalized line detector. On the 2-dimentional image plane, a line can be analytically described using parametric or normal notion as

$$\rho = x * cos(\theta) + y * sin(\theta) \quad (3)$$

where ρ is the length of a normal from the origin to this line and θ is the orientation of the norm with respect to the X-axis, refer to Fig.2 for example. Any point (x, y) on this line satisfies the definition:

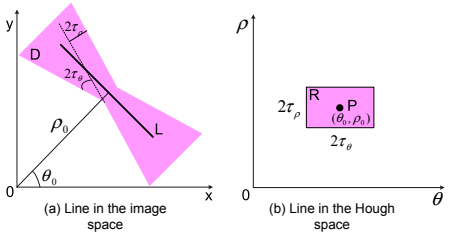

Figure 2. The correspondence between image space and Hough space: a point P in the Hough space will map to a line L in the image space; a region R in the Hough space will map to a partition D in the image space

$$\left\{(x,y) | \rho = F(x,y;\theta), (x,y) \in \chi^2\right\} \quad (4)$$

where $F(x,y;\theta) = x * cos(\theta) + y * sin(\theta)$ is the normal function of the line as in Equ.3, and χ^2 denotes the range of definition of the coordinate (x, y).

Theoretically, there is a one-to-one mapping between the lines in the image plane and the points in the Hough space which using ρ and θ as axis. Taking Fig.2 for example, a line L in image space corresponds to a point $P(\theta_0, \rho_0)$ in the Hough space. Therefore, we can use a point in the Hough space to represent a line in the image plane. We generalize this concept by extending a point in the Hough space to a rectangle region R which parameterized by the center position (θ_0, ρ_0) and window size $(2\tau_\theta, 2\tau_\rho)$. As illustrated in Fig.2(a), the back-projection of this region in the image plane is a butterfly-shaped region D. We refer to this region in the image plane as a *partition* to distinguish it from the rectangle region in the Hough space. Based on this extension, we can find a one-to-one mapping between a rectangle region in the Hough space and a partition in the image plane. This motivates us to generalize the definition of the line in Equ.4 as:

$$\left\{(x,y) | \rho = F(x,y;\theta), (x,y) \in \chi^2, |\rho - \rho_0| \le \tau_\rho, |\theta - \theta_0| \le \tau_\theta\right\} \quad (5)$$

We refer to this definition as *generalized line*. The geometry explanation of this definition is that all the line segments that fall into a rectangle region in the Hough space will be considered as a generalized line. This endows the generalized line with robustness to variations due to incontinuity, rotation and translation. More specifically, the robustness to rotation can be controlled quantitatively by parameter τ_θ and the robustness to translation can be controlled by τ_ρ, referring to Fig.2(a) for example. Therefore, these two parameters τ_θ and τ_ρ are referred to as *rotation uncertainty*

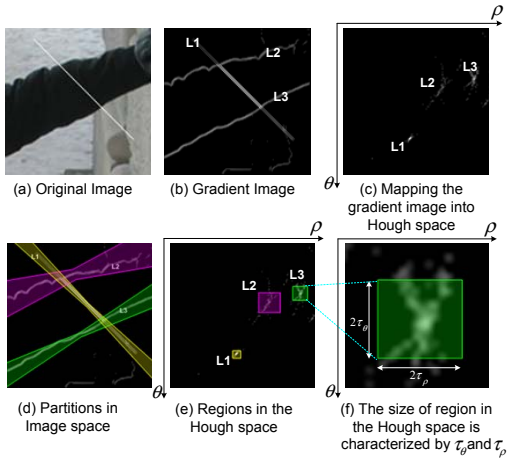

Figure 3. The lines with different linearity will map to the regions with different size in Hough space. Correspondingly, a region in Hough space will back-project to a partition in image space with certain level robustness to rotation, translation and incontinuity. This robustness can be controlled by the size parameters of the regions in Hough space

and *translation uncertainty* respectively. The standard definition of line in Equ.4 turns to a special case of the generalized line when τ_θ and τ_ρ are equal to zero.

The advantage of Equ.5 is that it can incorporate the uncertainty control into the line definition explicitly. Since for the applications in image processing and computer vision, we can seldom find a line that strictly meets the geometry definition in Equ.4 due to the imperfections in either the image data or the edge detector. Taking Fig.3 as an example, (a) is the original image (where white line in the image is actually artificial) and (b) is the edge image. Three "lines" can be observed in (b), denoted as L1, L2, and L3 respectively. Among them, only line $L1$ satisfies Equ.4 strictly, while the other two lines $L2$ and $L3$ are only approximate lines. Fig.3(c) visualizes the Hough space of (b), in which three clusters that correspond to the three lines in (b) can be observed. Here we can see that a "line" in the real-world image normally does not correspond to a single point in Hough space but a cluster of points instead. Therefore the generalized line can be more powerful for representing these imperfect lines. Moreover, the uncertainty in rotation and translation can be easily controlled by the region size (τ_θ, τ_ρ) in the Hough space. A smaller region corresponds to a generalized line with more specific geometry structure. Refer to Fig.3(d)∼(f) for example.

Since each generalized line has a certain level of uncertainty up to the limitations that specified by (τ_θ, τ_ρ), we can produce the lines with different description characteristics by varying the two uncertainty parameters. This property is just in line with the characteristic of granularity space that we've described in the Section 1. So the uncertainty parameters τ_θ and τ_ρ can be considered as a specialization of the granularity parameter τ. On the one hand, when the τ_θ and τ_ρ are set to some smaller values, the description of the line can be exact and deterministic, which is the basic property of the fine-granular features; on the one hand, when the τ_θ and τ_ρ are set to some bigger values, the description of the line can be statistical and robust to some variation, which is the desirable property of the coarse-granular features. Thus far, the feature parameter in Euq.1 can be specified as $\vartheta = (\theta_0, \rho_0)$ and granularity parameter can be specified as $\tau = (\tau_\theta, \tau_\rho)$. The image parsing function in Equ.2 can be represented as:

$$T(I; \vartheta, \tau) = T\left(I; \theta_0, \rho_0, \tau_\theta, \tau_\rho\right) \tag{6}$$

2.2. Image description

The image description function $S(\cdot)$ is to extract the descriptor of the generalized line that defined in Equ.5. The strength and the spatial distribution information are used for describing the generalized lines. More specifically, spatial distribution is represented by the normalized mean and the directional-variation of the positions of the gradients. It is worth noting that when the uncertainty factors τ_θ and τ_ρ are small, the variation along the norm direction does not take effect, since the shape variation has already been strictly constrained by the τ_θ and τ_ρ; when the uncertainty factors τ_θ and τ_ρ are larger, the shape constrain became weak and the specific shape of the line can be characterized by the variation alone the norm direction effectively. The detailed explanation of the feature extraction and normalization will be presented in the following section.

3. Implementation of GGP

According to the description in section 2.1, the definition of a generalized line is in the Hough space and can be considered as a $2\tau_\theta \times 2\tau_\rho$ window centered at (θ_0, ρ_0); but the description of this generalized line is in the image space. Therefore, we need to backproject the window from the Hough space into the image space. We achieve this goal by orientation partition and space partition.

3.1. Orientation-space partition

Orientation partition is to back-project the rotation uncertainty τ_θ into the image space. Given an image I, each pixel on the image can be represented as a triplet $[x, y, I(x, y)]$ where x and y denote the coordinates of the pixel and $I(x, y)$ denotes the intensity value. Then by applying the gradient operation, we can generate a gradient image dI, and each pixel on this gradient image can be represented by a quintet $[x, y, s, \theta, \rho]$, where $s = \sqrt{I_x^2 + I_y^2}$ represents the strength of gradient, $\theta = arctan(-I_x/I_y)$ denotes the tangent angle, and ρ can

be calculated as in Euq.3 (I_x and I_y denote the first-order derivatives of the intensity along the x and y directions). Then we quantized angle θ of the gradient image by step size τ_θ (rotation uncertainty) that has been defined in Equ.5. Therefore, we can divide the original gradient image into n disjoint orientation channels as $\{[x, y, s, \rho]_{\theta_0}, [x, y, s, \rho]_{\theta_1}, ..., [x, y, s, \rho]_{\theta_n}\}$, where $n = \lceil \pi/\tau_\theta \rceil$ and $\theta_i = i * \tau_\theta$. We refer to θ_i as the principle angle of each channel. For each channel $[x, y, s, \rho]_{\theta_i}$, only the pixels whose norm angle can be quantized as θ_i are preserved and all the other pixels' strength are set to zero. We refer to this operation as the orientation partition of the gradient image.

Space partition is to backproject the translation uncertainty τ_ρ into the image space. For each channel $[x, y, s, \rho]_{\theta_i}$ of the gradient image, we can further partition the channel image into parallel line belt regions, where the tangent angle of each partition line is equal to θ_i.

By the orientation and space partition, we can associate a window in the Hough space with a partition in the image space. Therefore, the statistical description of the generalized line can be calculated easily within this partition. And the granularity of the feature can be controlled easily by $\tau = (\tau_\theta, \tau_\rho)$ during the partition procedure. The overall orientation-space partition procedure can be seen in Fig.4.

3.2. GGP descriptor calculation and normalization

For orientation θ_i, GGP descriptor is a 7-dimensional heterogeneous vector, $(i_{max}, g_{max}, \sigma, m_x, m_y, v_{norm}, v_{tang})_{\theta_i}$. Given a feature specified by a rectangle R and the granularity parameter (τ_θ, τ_ρ), we firstly perform the orientation and space partition within the R as mentioned above. Then for each channel $[x, y, s, \rho]_{\theta_i}$, we can get the space partitions as $\{P_i | i = 1, \ldots, n; \bigcup_{i=1}^{n} P_i = R; \bigcap_{i=1}^{n} P_i = \phi\}$, where each P_i represents an individual partition and n is the number of partitions. The strength of the gradient on each partition can be calculated as $\{g_i | g_i = q(P_i); i = 1, \ldots, n\}$, where $q(\cdot)$ is the function that calculates the summation of strength within a partition. The items in the GGP descriptor can be calculated as follows:

- $i'_{max} = \arg\max_i (g_i)$ is the index of the region with the maximum gradient strength, and the normalized feature value can be calculated as $i_{max} = \frac{i'_{max}}{n}$
- $g'_{max} = \max(g_i)$ is the maximum gradient strength and can be normalized as $g_{max} = \frac{g'_{max}}{\sum_{i=1}^{n} g_i}$
- σ is the standard deviation of the gradient strength, and can be calculated as $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (g_i - \bar{g})^2}$, where $\bar{g} = \frac{1}{n} \sum_{i=1}^{n} g_i$
- m_x and m_y are normalized mean positions of all the non-zero pixels in partition $P_{i'_{max}}$, and can be calculated as $m_x = \frac{1}{t} \sum_{i=0}^{t} \frac{(x_i - x_0)}{w}$ and $m_y = \frac{1}{t} \sum_{i=0}^{t} \frac{(y_i - y_0)}{h}$, where t denotes the number of non-zero points, (x_0, y_0) represents the center of the rectangle R and (w, h) represents its size.
- v_{norm} and v_{tang} denote the standard deviations of positions of the non-zero pixels in partition $P_{i'_{max}}$ along the normal and tangent directions of the partition, which can be calculated as $v_{norm} = \sqrt{\frac{1}{t} \sum_{i=1}^{t} (r_{i,norm} - m_{norm})^2}$ and $v_{tang} = \sqrt{\frac{1}{t} \sum_{i=1}^{t} (r_{i,tang} - m_{tang})^2}$, where $\begin{bmatrix} m_{norm} \\ m_{tan} \end{bmatrix} = A \times \begin{bmatrix} m_x \\ m_y \end{bmatrix}$, $\begin{bmatrix} r_{i,norm} \\ r_{i,tan} \end{bmatrix} = A \times \begin{bmatrix} x_i \\ y_i \end{bmatrix}$, $A = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix}$ and θ_i is the principle angle of current channel.

Using this descriptor, we can address both the most prominent structure (generalized line) and overall gradient distribution. For example, the entries $(i_{max}, g_{max}, m_x, m_y, v_{norm}, v_{tang})_{\theta_i}$ are used for describing the partition that has the maximum gradient strength. Actually, this partition corresponds to the most prominent generalized line in direction θ_i. In addition, the entry $(\sigma)_{\theta_i}$ is used to capture the strength distribution information among all the partitions. The final feature vector is the concatenation of the GGP descriptors of all the orientation channels, and can be represented as $((i_{max}, g_{max}, m_x, m_y, v_{norm}, v_{tang})_{\theta_1}, \ldots, (\ldots)_{\theta_n})$.

3.3. Relation with HOG and edgelet

Since we have claimed that the GGP descriptor can cover different range in the granularity space, a question will be raised naturally: what are the two ends in this granularity space? Interestingly, we will show that the two ends of GGP correspond to two well-known features, HOG and edgelet.

A feature can be characterized by a rectangle $R(x, y, w, h)$ and granularity parameter (τ_θ, τ_ρ), where $\tau_\rho \in [1, \sqrt{w^2 + h^2}]$ and $\tau_\theta \in [0, \pi]$. Since τ_θ is used to control the number of quantized orientations, we only use τ_ρ to characterized the granularity setting.

For the coarsest granularity $\tau_\rho = \sqrt{w^2 + h^2}$, there will be only one space partition for each orientation channel, and the coverage of this partition is just the same as the rectangle $R(x, y, w, h)$. In this case, the GGP of each orientation becomes $(i_{max} = 0, g, \sigma = 0, m_x, m_y, v_{norm}, v_{tang})_{\theta_i}$ where g is the sum of gradient strength over the R. If we further discard spatial distribution information m_x, m_y, v_{norm} and

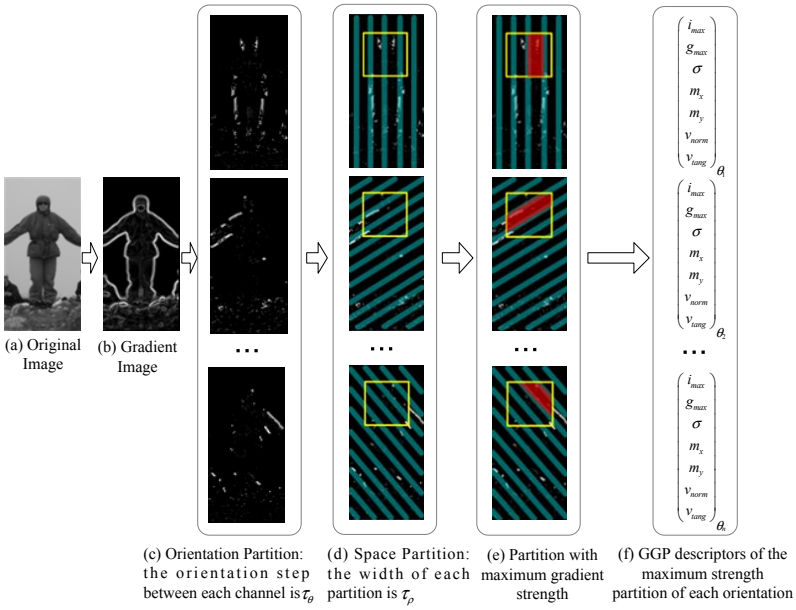

Figure 4. The overview of our approach: form the original image (a), a gradient image (b) is generated; (c) then we perform the orientation-partition to separate the gradient image into several channels according to different gradient directions, each channel is characterized by its principle orientation θ_i; (d) for each channel, we perform the space-partition to further divide each channel image into partitions as in (d) by a group of parallel lines which with the same orientation as θ_i; (e) within the feature window, the partition with the maximum gradient strength is selected for each orientation; (f) the GGP descriptor is calculated for each maximum gradient partition and concatenated together as a final descriptor

v_{tang}, the GGP will only contain one valid entry as $(g)_{\theta_i}$, and the final feature vector becomes $((g)_{\theta_1}, \ldots, (g)_{\theta_n})$, which is just a HOG descriptor with 1×1 cell.

For the finest granularity $\tau_\rho = 1$, the GGP becomes a line descriptor that can represent a line within the region R with the orientation θ_i, strength $(g_{max})_{\theta_i}$ and position $(m_x, m_y)_{\theta_i}$. In this case, GGP turns into a simplified edgelet that contains only one kind of basic shape: line segment.

4. Experiments

We evaluate the proposed method on INRIA [1] human dataset which contains the well-defined training and testing sets. It contains 1774 human samples (3548 with reflections) and 1671 human free images. For our evaluation, we follow the instruction of [1] and use the same training and testing data as [2].

The size of our normalized sample is 64×128. For training, all the 2416 positive training samples are cropped from the original 96×160 images with 16-pixel top margin and 16-pixel right margin. The 2436 background images (1218 background images and their reflections) are used as the pool for negative samples and bootstrapping. For testing, the 1132 positive samples are cropped from original 70×134 testing image with 3-pixel top margin and 3-pixel left margin. Following the instruction of [1], the negative testing samples are sampled from the 453 testing background images with 1.2 scale factor and 8-pixel scan step and the total number is 2141590. The detection error tradeoff (DET) curve on log-log coordinate is used to evaluate *classification* performance. The y-axis corresponds to the miss rate, $FalseNeg/(FalseNeg+TruePos)$, and the x-axis corresponds to false positives per window (FPPW), $FalsePos/(TrueNeg + FalsePos)$.

A LogitBoost classifier with rejection cascade is build for human detection. The weak classifiers are linear regression functions. For each stage of cascade, there are totally 2416 positive training samples and 10000 negative training samples. For the first stage, the negative samples are randomly selected from the background images; and for the following n_{th} stages, the negative samples are selected by bootstrapping of the previous $n-1$ stages. For each stage, the minimum detection rate is 99.7% and the maximum false positive rate is 35%. In addition, we also set a threshold of the margin between the separation boundary of the positive and negative samples, same as in [15].

For GGP feature, beside the position and size parameter (x, y, w, h), we have two more granularity parameters (τ_θ, τ_ρ). Even though we can use both τ_θ and τ_ρ to control the granularity for translation and rotation, in our implementation, we keep the $\tau_\theta = \pi/9$ as a fixed value and

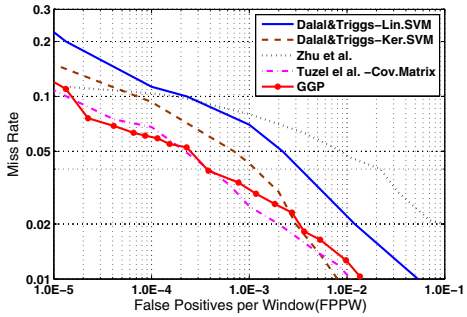

Figure 5. Comparisons with state-of-the-art methods on INRIA human dataset. The HOG descriptor with linear and kernel SVM from [2], HOG descriptor with cascade detector from [21] and covariance matrix methods from [15] are used as the baseline methods; GGP method can yield comparable results as covariance matrix method

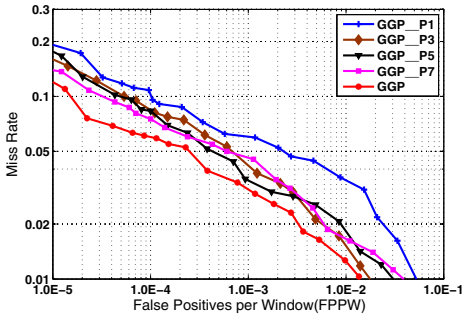

Figure 6. Evaluation of different granularity setting; the granularity parameters that can produce 1, 3, 5, 7 space partitions are evaluated; the combination of multiple granularity can yield best results

only use τ_ρ to control the granularity. By this means, the number of orientation channels will be fixed and we can use the integral image to calculate all the entries in the GGP descriptor efficiently. Further, we restrict the width and aspect ratio as $w \in \{16, 20, 24, 32, 38, 40, 50\}$ and $w/h \in \{1/0.5, 1/0.8, 1/1.0, 1/1.4, 1/1.8, 1/2.0, 1/2.2, 1/2.4\}$. The feature space is still very large even with these constrains. So for each stage, we randomly select 300 features as the feature pool for Adaboost feature selection.

In the first experiment, we choose the results from [2], [21] and [15] as the state of the art for benchmarking and a 26-stage cascade classifier is used for evaluation. By this classifier, we can achieve about 10^{-4} FPPW. For the last stage of bootstrapping, the scale factor is 1.05 and scan step is 3-pixel. Theoretically, we can train more stages by smaller scale factor and scan step, but after checking the bootstrapped negative samples, we found that most of 10000 negative samples come from just a few distinct patterns. This means the distribution of bootstrapped negative samples is biased and more training stages on current dataset can not guarantee better generalization ability. Similar problem has been observed in [15]. So we plot the points from stage 14 to 26 on the Fig.5 and then change the rejection threshold of the last stage to generate the points that $FPPW < 10^{-4}$. We use GGP to denote the results for this paper. From this figure we can see even we assume our feature in a linear space and use the linear function as the weak classifier, the combination of multiple-granularity feature can still yield better results than KernelSVM+HOG and comparable results as the covariance matrix.

In the second experiment, we intend to answer the question: what's performance if we use only one granularity? We test four translation uncertainties as $\tau_\rho \in \left\{ \lceil\sqrt{w^2+h^2}\rceil, \left\lceil\sqrt{\frac{w^2+h^2}{3}}\right\rceil, \left\lceil\sqrt{\frac{w^2+h^2}{5}}\right\rceil, \left\lceil\sqrt{\frac{w^2+h^2}{7}}\right\rceil \right\}$.

This means that within each feature window, we will test the features with 1, 3, 5 and 7 space partitions respectively. We use GGP_P1, GGP_P3, GGP_P5 and GGP_P7 to denote these four granularity settings and GGP to denote the combination of all the granularities. The experimental results can be seen in Fig.6. From these results, we can make a few observations:

- As we expected, the combination of multiple granularity can indeed yield the best results.
- On the DET figure, the effective regions of the different granularity features are different. For fine granularity features, better performance can be achieved when the FPPW is higher; for coarse granularity features, better performance can be observed when the FPPW is lower.
- The GGP_P1's performance is a bit worse than other granularities. This is out of our expectation. The possible reason is that among all the 7 entries in the GGP descriptor, there are two entries equal to zero for GGP_P1, referring to section 3.3 for example, which might reduce the discriminate ability of the feature.

In Fig.7, some detection results on INIRA set are presented. On the image, the thinner green boxes are the results of algorithm testing without post processing. The red boxes are the results of system testing with non-maximum supersession and overlapped-block merging. These results are achieved by applying the detector on the original image with 1.2 scale factor and 8-pixel scan step.

5. Conclusion

By introducing the new concept, *granularity space*, a family of descriptors with different representation ability is proposed to represent images/objects. Ranging from deterministic description to statistical representation, the pro-

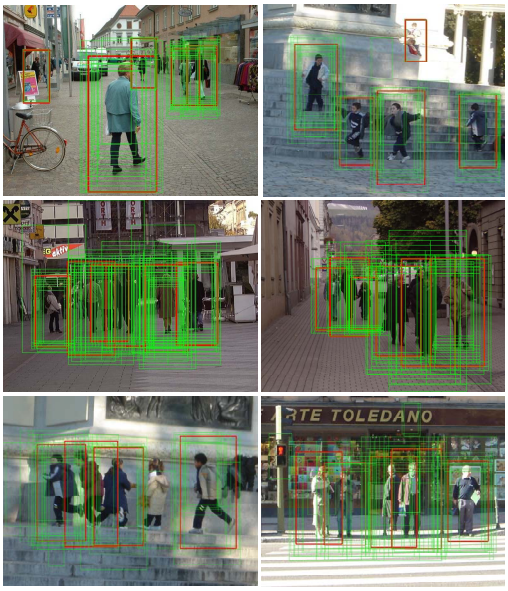

Figure 7. Some detection results on INRIA human dataset: the initial detection window is marked by the thin green rectangles; the final detection window with non-maximum supersession and overlap-window merging is marked by bold red rectangles; the scale factor is 1.2 and scan step is 8 pixel

posed descriptors can transit between the different representation properties easily by varying a single granularity parameter. This property can enrich the features' representation power significantly. We show that although the basic structure for GGP is very simple (only line segment), the descriptor is still capable of representing the complex objects, such as human body. Moreover, we also incorporate heterogeneous information into the descriptor, such as gradient's strength and spatial distribution information, which further improve the representation ability of the descriptor. Experiments on INRIA human dataset show that even we assume our features lie in a linear space and use linear weak classifier, the performance of the proposed GGP is still comparable to those of the-state-of-the-art methods that use either non-linear feature or non-linear classifier.

Acknowledgement

This paper is partially supported by NSFC under contracts Nos.60772071, 60833013, 60832004, 60872124; National Basic Research Program of China (973 Program) under contract 2009CB320902; Grand Program of International S&T Cooperation of Zhejiang Province S&T Department under contract No. 2008C14063; and Co-building Program of Beijing Municipal Education Commission.

The authors would also like to thank CherKeng Heng from Panasonic Singapore Laboratory for his helpful discussion.

References

[1] http://pascal.inrialpes.fr/data/human/. 6

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 1, 2, 6, 7

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441, 2006. 2

[4] P. Dollar, B. Babenko, S. Belongie, P. Perona, and T. Zhuowen. Multiple component learning for object detection. In *ECCV*, pages 211–224, 2008. 2

[5] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object detection by contour segment networks. In *ECCV*, pages 14–28, 2006. 2

[6] D. M. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV*, pages 37–49, 2000. 1

[7] F. Han, Y. Shan, H. S. Sawhney, and R. Kumar. Discovering class specific composite features through discriminative sampling with swendsen-wang cut. In *CVPR*, 2008. 2

[8] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR*, 2007. 2

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages 878 – 885, 2005. 2

[10] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *ICCV*, 2007. 1

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999. 2

[12] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, pages 69–81, 2004. 2

[13] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *TPAMI*, 23:349–361, 2001. 1

[14] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38:15–33, 2000. 1

[15] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007. 1, 2, 6, 7

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001. 1

[17] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741, 2003. 1

[18] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, pages 90 – 97, 2005. 2

[19] B. Wu and R. Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In *CVPR*, 2008. 2

[20] Y. Wu, C. Guo, and S. Zhu. From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, 2007. 2

[21] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, pages 1491– 1498, 2006. 2, 7